Rapport - Externe toetsingscommissie Samenspraak en Tegenspraak

16 januari 2015

Inleiding

Dit rapport beoordeelt het ontwerp en de werking van de tweede fase in de beoordeling van de partnerschapsaanvragen in het kader van het programma Samenspraak en Tegenspraak. In deze tweede fase gaat het om een inhoudelijke beoordeling van de aanvragen. Daaraan vooraf ging een beoordeling op zogenoemde drempelcriteria, die tot uitsluiting van een aantal van de ingediende voorstellen heeft geleid. Over deze eerste fase van de beoordelingsprocedure heeft de commissie in een eerder stadium gerapporteerd.

De taak van de externe toetsingscommissie is om de selectie van maatschappelijke partners van het ministerie in het kader van het programma Samenspraak en Tegenspraak te beoordelen op kwaliteit, consistentie en zorgvuldigheid. De commissie is ingesteld als externe, onafhankelijk 'check' op het proces.

De commissie baseert haar rapport op een aantal bronnen:
- een serie verantwoordingsrapportages van het plv hoofd van de afdeling Maatschappelijke Organisaties (DSO/MO), J. van Bommel;
- een verantwoordingsrapportage door het Vliegende Keep Team;
- inzage in 15 aselect gekozen aanvraagdossiers en daarop uitgevoerde beoordelingen door het relevante beoordelingsteam, het vliegende keep team en de beoordelingscommissie;
- vergelijkende tabellen en overzichten vervaardigd door het vliegende keep team respectievelijk de afdeling Maatschappelijke Organisaties, deels specifiek op verzoek van de commissie;
- vraaggesprekken met hoofd en plv hoofd van de afdeling MO, een lid van het Vliegende Keep Team, en twee leden van de beoordelingscommissie, en daarvan gemaakte gespreksverslagen.

Dit tweede rapport is als volgt opgebouwd: eerst schetsen en beoordelen wij de achtergrond en het ontwerp van het beoordelingssysteem. Daarna gebeurt dat voor de feitelijke werking van het systeem. Tenslotte beoordelen wij het product van het systeem, in casu het voorstel van de beoordelingscommissie. Tot slot doen wij een aantal aanbevelingen voor het ontwerp en management van toekomstige, vergelijkbare procedures.

Doelen en ontwerp van het beoordelingssysteem

Beschrijving
Oogmerk van de hele operatie is om tot selectie van maximaal 25 maatschappelijke organisaties te komen die als partner van het ministerie van fungeren in de uitvoering van het beleidskader Samenspraak en Tegenspraak. Centraal daarin staat de ambitie om te gaan werken met organisaties/allianties

die lokale partners in het globale Zuiden in staat stellen om effectief te 'pleiten en beïnvloeden' op een aantal beleidsmatig prioritaire thema's. Kandidaat-partners is gevraagd in hun aanvraag in het bijzonder uiteen te zetten:
a) wat de zogenoemde 'theory of change' is van waaruit zij hun activiteiten op het gebied van capaciteitsversterking in pleiten en beïnvloeden vormgeven
b) wat hun zogenoemde 'track record' op dit terrein is.

Na de uitvoering van de drempeltoets resteerden 56 organisaties/consortia wier aanvragen in aanmerking kwamen voor inhoudelijke beoordeling. Het beoordelingssysteem kende de volgende componenten/stappen:

- De indiening van de aanvraag, met daarin in het bijzonder de Track Record (TR), de Theory of Change (ToC)
- De instelling en training van ambtelijke beoordelingsteams bestaande uit drie personen (twee ambtelijk, waarvan een afkomstig uit DSO/MO en een met relevante thematische expertise, en een externe deskundige) die aan de hand van een centraal vervaardigd beoordelingsstramien de verschillende componenten van TR en ToC van scores voorzien, op een vijfpuntsschaal (A-E). Onderdeel van deze fase van de beoordeling zijn voorts: a. een korte mondelinge 'pitch' door de aanvrager ten overstaan van het relevante beoordelingsteam; b. het opvragen van opinies van in de aanvraag genoemde externe referenten (maximaal drie). Het resultaat van de beoordeling wordt vastgelegd in 'five pagers': ingevulde beoordelingsformulieren voor zowel TR als ToC, waarin de gegeven scores worden beargumenteerd in het licht van de aangereikte beoordelingscriteria.
- Toetsing op interne en vergelijkende consistentie van de door de beoordelingsteams (BT's) uitgevoerde beoordeling door een centraal 'vliegende keep team' (VKT). Het VKT traint ook de BT's in het werken met het beoordelingsstramien en is tussentijds beschikbaar voor advies aan en 'meekijken' met de BT's. In geval van verschillende taxaties van een bepaalde aanvraag door BT en VKT volgt overleg/afstemming.
- De initiële BT-beoordelingen, de second opinions van het VKT en de eventueel bijgestelde BT-beoordelingen worden doorgeleid naar de centrale beoordelingscommissie (BC). Daarbij worden een vooraf vastgestelde volgorde en tijdschema aangehouden
- De BC krijgt de beschikking over de aanvraag, de five-pager en het oordeel van het VKT. Zij voert daarnaast vraaggesprekken met alle aanvragers, mede gevoed door adviezen vanuit het BT om bijzondere aandacht te besteden aan bepaalde aspecten van de aanvraag. Deze bijzondere aandachtspunten worden tevoren schriftelijk medegedeeld aan de aanvrager. De commissie kan daarnaast om toelichting verzoeken bij VKT en BT's, respectievelijk thema-experts vanuit het departement.
- De BC stelt haar eindoordeel ook in termen van een A-E schaal vast. Het oordeel wordt gemotiveerd in een zgn 'one pager', ook wel fiche genoemd.
- Tot stelt de commissie een integrale ranking op en formuleert op basis daarvan een advies aan de minister.

Beoordeling
De commissie is onder de indruk van de ambitie van het ministerie om de selectie van partners te laten plaatsvinden op een manier die hen maximaal de ruimte biedt zichzelf inhoudelijk te presenteren, zowel schriftelijk als verbaal. Nadruk in de hele beoordelingsprocedure ligt op inhoudelijkheid en argumentatie. De commissie acht dit een verfrissend contrast met de bij de rijksoverheid veel gebruikelijker 'kadaverdiscipline' die aan aanvragers wordt opgelegd in het kader van gestandaardiseerde aanbestedingsprocedures: weinig ruimte voor inhoud, veel nadruk op makkelijker te standaardiseren en kwantificeren governance- en bedrijfsvoeringsfactoren.

Er zit aan die ambitie ook een prijs. Van de *aanvragers* wordt veel, heel veel gevergd. De 'opportunity costs' van het werken aan de aanvraag zijn omvangrijk. Daar staat tegenover dat de toekenning van een partnerschap grote financiële en relationele verplichtingen met zich mee brengt voor beide partijen. Er mag dus ook best wat worden gevergd. Van het *ministerie* vergde het ontwerpen en uitvoeren van een toetsingssysteem waarin inhoudelijke rijkheid optimaal recht wordt gedaan en tegelijkertijd tot een sluitende, verdedigbare selectie van maximaal 25 partners moet worden gekomen, een enorme krachtsinspanning. De bovenstaande schets van de procedure geeft een indruk van die inspanning. Het is goed dat nog expliciet te memoreren: opstelling van het toetsingskader (inzet van externe deskundigen); 65, later 56 aanvragen (zeer omvangrijke dossiers), 168 referenties die moesten worden opgevraagd, acht beoordelingsteams van 3 personen (inclusief 1 externe per team), een vliegende keep team (inclusief 1 externe), een gemengd intern/externe beoordelingscommissie, 56 interviews, etc. Wat is de kosten/baten-verhouding van al deze inspanningen?

De commissie meent dat de balans per saldo duidelijk positief uitvalt. Het omvangrijke systeem van kwalitatieve toetsing door drie verschillende gremia (BT's, VKT en uiteindelijk BC) zorgt ervoor dat niet alleen bij de aanvragers maar ook in de beoordeling het primaat van inhoud en argumentatie geldt. Het getrapte karakter maakt dat het checks and balances tussen beoordelaars bevat. Dit zorgt ervoor dat het uiteindelijke oordeel intersubjectief gevalideerd is. Het beoordelingsproces is bovendien transparant en dus navolgbaar in de vorm van de 5-pagers (van de BT's) en de 1-pagers (van de BC). De positionering van het VKT als vraagbaak en second opinion brengt bovendien een extra element van dialoog en correctie in.

Dat betekent niet dat het systeem in zijn feitelijke werking geheel vlekkeloos is gebleken. Hierop gaan wij in de volgende paragraaf in.

Het systeem in de praktijk

Beschrijving
De feitelijke werking van het bovenbeschreven systeem is op een aantal manieren afgeweken van het ontworpen ideaal.

Er was in het hele proces sprake van tijdsdruk: voor een deel werd er nog aan het systeem gesleuteld terwijl het al in werking moest treden. De aanvragen kwamen in september binnen. Er waren aanvragers strakke termijnen toegezegd voor uitvoering van de drempeltoets en bekendmaking van de resultaten van de inhoudelijke toetsing. Maar in de 'back office' moest ondertussen nog van alles worden geregeld. Samenstelling van de teams was een uitdaging, omdat er veel overwegingen recht moest worden gedaan: onpartijdigheid, evenwichtigheid, relevante expertise, beschikbaarheid, timing. Veel tijd/gelegenheid om alle teamleden systematisch en uniform te trainen in het hanteren van de ontwikkelde beoordelingsstramienen was er uiteindelijk niet.

Mede als gevolg daarvan was het niet mogelijk om volgtijdelijk te werk te gaan. Idealiter zou het werk van alle BT's zijn afgerond en nog eens goed zijn vergeleken en gecorrigeerd voordat de BC zou optreden. Dat is onmogelijk gebleken. Er werd dus in belangrijke mate parallel gewerkt, hetgeen betekent dat de BC soms niet-voldragen input van de BT's kreeg. Wellicht mede daarom heeft het de BC zich tamelijk autonoom opgesteld ten opzichte van het ambtelijk voorwerk in BT's/VKT. Hier staat tegenover dat BC niet 'voortgeborduurd' heeft op de mening van de BT's, maar veelal eerst een eigen mening gevormd had, alvorens met het BT/VKT in discussie te gaan, wat wellicht geleid heeft tot een grotere rijkheid aan overwegingen.

Geconfronteerd met de beperkte mogelijkheden van een vijfpuntsschaal om nuances in het oordeel aan te brengen, heeft VKT besloten gebruik te maken van +/- kwalificaties voor aanvragen die in de C en D-categorieën leken te vallen. De BC heeft dat niet gedaan, maar heeft desalniettemin wel geconcludeerd dat sommige C-aanvragen 'dikke' en andere 'normale' of 'dunne' C-s waren. Daarmee is wel enige ambiguïteit geslopen in de betekenis van deze (sub)kwalificaties, die vervolgens door onderlinge discussies weer moest worden opgehelderd/onderhandeld.

Tijdens het werken met het beoordelingsstramien voor ToC en TR werd duidelijk dat er interpretatie- en beoordelingsverschillen mogelijk waren, en ook feitelijk optraden, binnen en tussen BT's. Dit noopte tot intensieve discussies en bemiddelend werk door het VKT. De BC bleek slecht uit de voeten te kunnen met de beoordelingsstramienen door het ontbreken van een eenduidige aggregratie- en wegingssystematiek binnen de ToC en TR beoordelingen. Ook ontbrak het aan een eenduidig regel voor het relatieve gewicht van TR- en ToC-scores in de totstandkoming van het eindoordeel.

De relatieve meerwaarde van de 'pitches' en het opvragen van referenties bleek tamelijk beperkt. Referenties waren niet altijd makkelijk te krijgen, en bovendien ook weinig discriminerend (de referenten waren vermoedelijk strategisch gekozen). De pitches waren erg kort en uitsluitend 'eenrichtingsverkeer'. Ze dwongen weliswaar de aanvragers om door te dringen tot de echte kern van hun voorstel, maar de BT's was niet duidelijk gemaakt hoe ze deze pitches nu precies bij hun oordeelsvorming moesten betrekken.

De interviews met de aanvragers door de BC werden daarentegen als buitengewoon nuttig ervaren: de aanvragers konden hun ei kwijt, nadere toelichting verschaffen, en een gevoel krijgen hoe hun schriftelijke aanvraag was gelezen door de commissie. De commissie kon verheldering krijgen, aanvragers de nieren proeven op punten waarin de aanvraag niet duidelijk of niet overtuigend was geweest. Ook hier gold evenwel dat niet tevoren 100% duidelijk was of en hoe het gesprek zou worden meegewogen in de beoordeling. De commissie zegt hier zelf over dat informatie die niet expliciet terug te vinden was in de schriftelijke aanvraag ook niet is meegenomen in de weging.

Beoordeling
Hoewel het proces een aantal zwakheden in de beoordelingssystematiek heeft blootgelegd, zijn die voldoende gecompenseerd door de doelbewuste redundantie (BT, VKT en BC) en interactie (overleg, advisering, interviews) en het gebruik van aanvullende expertise (bv over gender/diversiteit) die in het proces is ingebouwd.

Idealiter is een systeem van meerdere beoordelaars dermate strak en eenduidig dat er grote correspondentie is tussen hun oordelen. De mate van correspondentie in het onderhavige systeem was als volgt: op 21 van de 56 oorspronkelijke BT-beoordelingen kwam het VKT tot een ander oordeel dan het betreffende BT (in 13 gevallen in neerwaartse richting, overigens vaak doordat het VKT met -/+ toevoegingen nuances ging aanbrengen waarover de BT's niet beschikten); in 11 van de 56 gevallen kwam het BC tot een ander eindoordeel dan het gezamenlijke BT/VKT-oordeel (overigens deels omdat de BC behoudens in een geval geen gebruik maakte van het +/- systeem van het VKT) .

Gelet op het sterk kwalitatieve, interpretatieve karakter van de procedure is dat geen ideaal, maar zeker ook geen slecht resultaat. De *mate* van afwijking tussen beoordelaars was bovendien relatief beperkt: het ging zoals gezegd vaak om +/- nuances niet om 'volle' letterverschillen in de taxaties van ToC's en TR's.

Desalniettemin constateert de commissie dat het systeem in werking een aantal problematische aspecten vertoonde:

1. De beperkte operationalisatie en daardoor de gebrekkige eenduidigheid van de beoordelingscriteria voor de verschillende deelaspecten van TR en ToC die in het beoordelingsstramien waren onderscheiden. Per deelcriterium zijn duidelijker en systematischer volgehouden 'schalen' van indicatoren nodig om tot heldere, transitieve A-E oordelen te komen.
2. Het ontbreken van een eenduidige aggregratie- en wegingssystematiek binnen de ToC en TR beoordelingen, alsmede in het relatieve gewicht van TR respectievelijk ToC in de totstandkoming van het eindoordeel.
3. De ad-hoc invoering van +/- nuanceringen door de VKT veranderende wat eerst een 5-puntswaardering was, de facto, in een 9-puntswaardering (A, B, C+, C, C-, D+, D, D-, E). De BC maakte geen gebruik van dat systeem, maar zocht vervolgens toch ook weer naar een weg om differentiatie binnen de C-stapel te kunnen aanbrengen. Dat is meer gevoelsmatig/argumenterend dan systematisch/coderend tot stand

gekomen, en vergroot daarmee de kwetsbaarheid van de uiteindelijke oordeelsvorming.

4. De uiteenlopende kwaliteit van de argumentaties in de invulling van de beoordelingsstramienen door de BTs. Dit is ook het VKT en de BC opgevallen en er is werkenderweg zo veel mogelijk geprobeerd evenwicht te creëren en de gemiddelde kwaliteit te verhogen.
5. Het feit dat de 5-pagers van de BT's en de 1-pagers (fiches) van de BC's nu co-existeren. De BC heeft weliswaar gebruik gemaakt van de 5-pagers bij het schrijven van haar fiches, maar selectief en met behoud van eigen oordeelsvorming. De fiches vormen straks de belangrijkste input bij het schrijven van de toekennings- en afwijzingsbeschikkingen. Het is (vermoedelijk ook vanuit juridisch oogpunt) kwetsbaar dat er onder/achter die fiches dus (opvraagbare) 5-pagers liggen waarin soms tot andere beoordelingen wordt geconcludeerd en andere redenaties worden gehanteerd.

Al met al heeft de commissie voldoende vertrouwen in de integriteit van het proces om te concluderen dat de A en B-beoordelingen boven gerede twijfel verheven zijn. Ook de toekenning van C's is als zodanig vermoedelijk goed verdedigbaar, zo lang duidelijk is dat de uiteindelijke toekenning een exclusieve taak van de BC is, 'gehoord' maar niet 'in overeenstemming met' de BT's of het VKT. In die interpretatie van het systeem is dan ook de motivatie in de 1-page fiches geldig; de 'onderliggende' 5-pagers van de BT's en de opinies van het VKT zijn daarvoor dan een ingrediënt, en gelden niet als zelfstandig oordeel waaraan partijen rechten kunnen ontlenen.

Het advies van de beoordelingscommissie

Beschrijving
De BC heeft uiteindelijk een advies geformuleerd waarin niet alleen een ranking van de aanvragers op het A-E continuum is neergelegd, maar ook een aantal scenario's is aangereikt voor de minister om tot 25 partners te komen. In het advies werden 18 aanvragen met A en B beoordeeld, en dus als preferente partners voorgesteld. Daarnaast werden drie van de 28 aanvragen die met een C zijn beoordeeld voorgesteld als de 'top drie' van de C's, die ook voor partnerschap in aanmerking zouden moeten komen. Gelet op het feit dat er maximaal 25 partnerschappen beschikbaar zijn, heeft de commissie vervolgens twee 'scenario's' aangereikt om op basis van bepaalde beleidsprioriteiten nog eens vier van de resterende C-aanvragen een partnerschap aan te bieden.

Beoordeling
De beoordelingscommissie heeft haar mandaat ruim opgevat. Zij heeft niet alleen een 'beoordeling' (ranking) van de aanvragen opgeleverd, maar heeft zich ook opgesteld als adviseur van minister. Daarbij heeft zij het op zich genomen om – al dan niet op instigatie van het departement - een oplossing aan te reiken voor de spanning die ontstond tussen:

a. Het feit dat er 18 onomstotelijk honoreerbare A/B-aanvragen uit de beoordelingsprocedure naar voren kwamen en 28 aanvragen een (mogelijk ook honoreerbare) C-status kregen;

b. de kennelijk op het ministerie levende interpretatie dat de 'maximaal' 25 beschikbare partnerschappen ook allemaal daadwerkelijk moeten worden gehonoreerd in de huidige procedure.

De BC is daarom vervolgens op twee punten gaan sleutelen:

- haar gevoelsmatige oordeel dat er in de C-stapel drie heel 'dikke C's' waren, die duidelijk beter waren dan de rest is omgezet in een advies om ook deze C's een partnerschap aan te bieden
- het aanbieden van twee scenario's (innovatie, diversiteit) die ieder op hun beurt vier van de resterende C-aanvragen prviligieerden.

De toetsingscommissie begrijpt de wens van de BC om in de opstelling van haar advies constructief 'mee te denken' (waartoe zij wellicht ook expliciet is uitgenodigd), maar meent dat het onwenselijk is vanuit het oogpunt van de zuiverheid en robuustheid van de beoordelingsprocedure. Het geformuleerde advies is naar onze inschatting kwetsbaar voor beroep. Het is ook niet nodig: in de procedure is ingebouwd dat de minister discretionaire ruimte heeft en deze op grond van beleidsmatige overwegingen mag invullen. De beoordelingsprocedure biedt de minister een stevig fundament om 18 partnerschappen direct in te vullen. Zij biedt als zodanig echter geen eenduidig evaluatief fundament om vervolgens nog eens maximaal zeven additionele partners te kiezen. Voor zover de minister dat toch wil, zal zij dat niet op grond van kwalitatieve maar op grond van beleidsmatige (bv 'goodness of fit' met door haar te prioriteren beleidsthema's) moeten doen. Advisering daarover (zoals in de twee scenario's van de BC) is een taak van de ambtelijke organisatie, niet van de BC, tenzij de Minister dat expliciet en gericht aan de commissie verzoekt.

Balans

1. De commissie heeft grote waardering voor de ambitie en uitgangspunten van het beoordelingssysteem voor partnerselectie binnen Samenspraak en Tegenspraak. Het is gedurfd en innovatief. Het is juist daarom echter ook complex en potentieel kwetsbaar in zijn uitvoering.
2. De grote inzet, integriteit en professionaliteit van de ambtelijke organisatie (en de beoordelingscommissie) in het proces verdienen niets dan lof. Men heeft in de geest van het programma gepionierd en geprobeerd een zeer complex, arbeidsintensief en nog niet getest systeem te laten werken, en dat alles onder stevige tijdsdruk. Daarbij zijn de nodige improvisaties toegepast, die over het algemeen goed verdedigbaar zijn en waarover de organisatie ook transparant is geweest ten opzichte van deze commissie. Desalniettemin heeft het in werking brengen van het systeem pas goed duidelijk gemaakt waar de voetangels en klemmen zitten.
3. Het resultaat van alle inspanningen is dat er 18 partners zonder meer kunnen worden aangewezen. De keuze voor deze partijen berust op een zeer hoge mate van convergentie in het oordeel van BT's, VKT en BC.
4. Het resultaat is evenwel ook dat er op dit moment geen waterdichte basis ligt om dit aantal op te hogen tot (om en nabij) 25. Daarvoor zal de minister ofwel een aanvullende, lichte procedure in het leven moeten

roepen die voortborduurt op de al verrichte inspanningen, ofwel gebruik moeten maken van haar discretionaire ruimte om maximaal zeven van de 28 organisaties die kwalitatief gezien vergelijkbaar (namelijk: C-score) uit de bus zijn gekomen, alsnog een partnerschap aan te bieden op grond van beleidsmatige prioriteiten.

Prof dr P. ‘t Hart (voorzitter), drs. J. Jeurissen (lid), drs. H. Jorritsma (lid), m.m.v. J. Hart (ambtelijk secretaris)